International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 6, Nov-Dec-2020, ISSN (Online): 2395-566X

Deep Semantic Image Segmentation using Convolutional Neural Networks for Multi Modal Data

Asst.Prof. P Jaganmohan , Asst. Prof. R Suneel Kumar
Dadi Institute of Engineering and Technology
Anakapalle, Visakhapatnam, A.P.
pjmohan@diet.edu.in,suneelkumar@diet.edu.in

***Abstract*- In this paper, a deep semantic segmentation of aerial imagery based on multi-modal data is discussed. given multi-modal data composed of true orthophotos and the corresponding digital surface models (dsms), we extract a variety of handcrafted radiometric and geometric features which are provided separately and in different combinations as input to a modern deep learning framework. the latter is represented by a residual shuffling convolutional neural network (rscnn) combining the characteristics of a residual network with the advantages of atrous convolution and a shuffling operator to achieve a dense semantic labeling. via performance evaluation on a benchmark dataset, we analyze the value of different feature sets for the semantic segmentation task. the derived results reveal that the use of radiometric features yields better classification results than the use of geometric features for the considered dataset. furthermore, the consideration of data on both modalities leads to an improvement of the classification results. however, the derived results also indicate that the use of all defined features is less favorable than the use of selected features. consequently, data representations derived via feature extraction and feature selection techniques still provide a gain if used as the basis for deep semantic segmentation.**

Keywords- semantic segmentation, multi-modal data, deep learning, cnn, residual network

I. INTRODUCTION

The semantic segmentation of aerial imagery in terms of assigning a semantic label to each pixel and thereby providing meaningful segments has been addressed in the scope of many recent investigations and applications. While the radiometric information preserved in an orthophoto can already be sufficient to distinguish specific classes, the geometric information preserved in the corresponding DSM might alleviate the separation of further classes, as each modality provides information about different aspects of the environment.

Generally, the semantic segmentation of aerial imagery based on true orthophotos and the corresponding DSMs can be achieved via the extraction of handcrafted features. Nowadays, however, many investigations rely on the use of modern deep learning techniques. Some of these approaches also focus on using hand-crafted features derived from the true orthophotos or from their correspond- ing DSMs in addition to the given data as input to a deep learning technique. In this regard, the Normalized Difference Vegetation Index (NDVI) and the normalized Digital Surface Model (nDSM) are commonly used. Other kinds of hand-crafted features have however only rarely been involved so far although they might introduce valuable information for the semantic labeling task. In this paper, we focus on the deep semantic segmentation of aerial imagery based on multi-modal Data. We extract a diversity of handcrafted features from both the true orthophotos and their corresponding DSMs. Based on a separate and combined consideration of these radiometric and geometric features, we perform a supervised classification involving modern deep learning techniques. As standard deep networks (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014) are composed of many layers to learn complex non-linear relationships, such networks tend to suffer from the vanishing gradient problem if they are very deep, i.e. the gradients backpropagated through the layers become very small so that the weights in early layers of the network are hardly changed.

This, in turn, causes a decrease in the predictive accuracy of the network and can be resolved by using a Residual Network (ResNet) (He et al., 2016a). Relying on the ResNet architecture originally intended to classify image patches, we present a modified ResNet architecture that allows a dense semantic image labeling.

More specifically, we make use of the ResNet-34 ar chitecture and introduce both atrous convolution and a shuffling operator to achieve a semantic labeling for each pixel of the input imagery. We denote the resulting deep network as Residual Shuffling Convolutional Neural Network (RSCNN). Via performance evaluation on a benchmark dataset, we quantify the effect of considering the different modalities separately and in combination as input to the RSCNN. We observe that the

additional extraction of different types of geometric features based on the DSM and the definition of corresponding feature maps for the RSCNN leads to an improvement of the classification results, and that the best classification results are achieved when using selected feature maps and not when using all defined feature maps.

II. RELATED WORK

For many years, the semantic segmentation of aerial imagery based on multi-modal data has typically been addressed by extracting a set of handcrafted features (Gerke and Xiao, 2014; Tokarczyk et al., 2015; Weinmann and Weinmann, 2018) and providing them as input to a standard classifier such as a Random Forest (Weinmann and Weinmann, 2018) or a Conditional Random Field (CRF) (Gerke, 2014). Due to the great success of modern deep learning techniques in the form of Convolutional Neural Networks (CNNs), however, many investigations nowadays focus on the use of such techniques for semantically segmenting aerial imagery as they tend to significantly improve the classification results.

Regarding semantic image segmentation, the most popular deep learning techniques are represented by Fully Convolutional Networks (FCNs) (Long et al., 2015; Sherrah, 2016) and encoder- decoder architectures (Volpi and Tuia, 2017; Badrinarayanan et al., 2017). The latter are composed of an encoder part which serves for the extraction of multi-scale features and a decoder part which serves for the recovery of object details and the spatial dimension and thus addresses a more accurate boundary localization. A meanwhile commonly used encoder-decoder structure has been proposed with the SegNet (Badrinarayanan et al., 2017).

Specifically addressing semantic segmentation based on multi-modal data in the form of orthophotos and the corresponding DSMs, different strategies to fuse the multi-modal geospatial data within such a deep learning framework have been presented (Marmanis et al., 2016; Audebert et al., 2016; Audebert et al., 2017; Liu et al., 2017), while the consideration of semantically meaningful boundaries in the SegNet encoder-decoder architecture and also in FCN-type models has been addressed by including an explicit object boundary detector to better retain the boundaries between objects in the classification results (Marma- nis et al., 2018). As an alternative to involving a boundary detec- tor, it has been proposed to discard fully-connected layers (which reduce localization accuracy at object boundaries) and to additionally avoid the use of unpooling layers (which are more complicated and e.g. used in SegNet) (Chen et al., 2017). In this paper, we investigate the value of different types of handcrafted features for the semantic segmentation of aerial imagery based on multi-modal data. We extract a diversity of hand-crafted features from both the true orthophotos and their corresponding DSMs. Thereby, we involve hand-crafted radiometric features such as the NDVI and one of its variants, but also radiometric features derived from transformations in analogy to the definition of color invariants (Gevers and Smeulders, 1999). Furthermore, we involve hand-crafted geometric features in the form of the nDSM (Gerke, 2014) and features extracted from the 3D structure tensor and its eigenvalues.

III. METHODOLOGY

The proposed methodology addresses the semantic interpretation of aerial imagery by exploiting data of several modalities (Section 3.1) which are provided as input to a deep network (Section 3.2). The result is a dense labeling, i.e. each pixel is assigned a respective semantic label.

1. Feature Extraction

Given a true orthophoto and the corresponding DSM on a regular grid, the information may be stored in the form of a stack of feature maps (i.e. images containing the values of a respective feature on a per pixel level), whereby three feature maps correspond to the spectral bands used for the orthophoto and one fea- ture map corresponds to the DSM. Further information can easily be taken into account by adding respective feature maps. In total, we define eight radiometric features (Section 3.1.1) and eight geometric features (Section 3.1.2) for the given regular grid. Based on these features, we define corresponding feature maps which serve as input to a CNN.

2. Radiometric Features

In our work, we assume that the spectral bands used for the orthophoto comprise the near-infrared (NIR), red (R) and green (G) bands (Cramer, 2010; Rottensteiner et al., 2012; Gerke, 2014). Accordingly, we define the reflectance in the near-infrared domain, in the red domain and in the green domain as features denoted by the variables R_{NIR}, R_R and R_G, respectively. In addition, we consider color invariants as features. In analogy to the definition of color invariants derived from RGB imagery to improve robustness with respect to changes in llumination, we consider normalized colors which represent a simple example of such color invariants

3. Geometric Features

In addition to the radiometric features, we extract a set of geometric features. The most intuitive idea in this regard is to take into account that the heights of objects above ground are more informative than the DSM itself. Consequently, we use the DSM to calculate the normalized Digital Surface Model (nDSM) via the approach presented in (Gerke, 2014). This approach relies on first classifying pixels into ground and off-ground pixels using the LAStools software. Subsequently, the height of each off-ground pixel is adapted by subtracting

the height of the closest ground point. Besides the nDSM, we involve a set of local shape features extracted from the DSM as geometric features.

4. Supervised Classification

For classification, we focus on the use of modern deep learning techniques in the form of convolutional neural networks, where standard networks like AlexNet (Krizhevsky et al., 2012) and the VGG networks (Simonyan and Zisserman, 2014) are composed of a collection of convolutional layers, max-pooling layers and activation layers followed by fully-connected classification layers. The use of deep networks with many layers allows learning complex non-linear relationships, yet it has been found that the performance of very deep networks tends to decrease when adding further layers via simply stacking convolutional layers.

5. Atrous Convolution

As the field-of-view of the deeper layers will shrink after removing RRLs, we involve atrous convolution (Chen et al., 2016) which can be used to compute the final CNN responses at an arbitrary resolution through re-purposing the networks trained on image classification to semantic segmentation and to enlarge the field-of-view of filters without the need for learning any extra parameters.

6. Shuffling Operator

To achieve a dense prediction, we involve a shuffling operator to increase the resolution by combining feature maps in a periodic shuffling manner. The concept of the shuffling operator has been originally introduced for super resolution (Shi et al., 2016) and it aims at the upscaling of feature maps. Inspired by this idea, it has been proposed to introduce this operator for the semantic segmentation of aerial imagery (Chen et al., 2018), and respective experiments reveal that the use of a shuffling operator improves the predictive accuracy through forcing networks to learn upscaling.

IV. EXPERIMENTAL RESULTS

In the following, we first describe the used dataset (Section 4.1). Subsequently, we summarize the conducted experiments (Section 4.2) and, finally, we present the derived results (Section 4.3).

1. Dataset

For our experiments, we use the Vaihingen Dataset (Cramer, 2010; Rottensteiner et al., 2012) which was acquired over a relatively small village with many detached buildings and small multi-story buildings. This dataset contains 33 patches of different sizes, whereby the given regular grid corresponds to a ground sampling distance of 9 cm. For 16 patches, a very high-resolution true orthophoto and the corresponding DSM derived via dense image matching techniques are provided as well as a referencelabeling with respect to six semantic classes represented by Impervious Surfaces, Building, Low Vegetation, Tree, Car and Clutter/ Background. According to the specifications, the class Clutter/ Background includes water bodies and other objects such as containers, tennis courts or swimming pools. We use 11 of the labeled patches for training and the remaining 5 labeled patches for evaluation.

4.2 Experiments

For each orthophoto and the corresponding DSM, we extract the set of hand-crafted features (cf. Section 3.1). Based on the orthophoto, we derive eight feature maps containing radiometric information with respect to the reflectance in the near-infrared (NIR), red (R) and green (G) domains, the normalized nearinfrared (nNIR), normalized red (nR) and normalized green (nG) values, the Normalized Difference Vegetation Index (NDVI) and the Green Normalized Difference Vegetation Index (GNDVI). Based on the DSM, we derive eight feature maps containing geometric information with respect to the normalized Digital Surface Model (nDSM), linearity (L), planarity (P), sphericity (S), omnivariance (O), anisotropy (A), eigenentropy (E) and change of curvature (C).

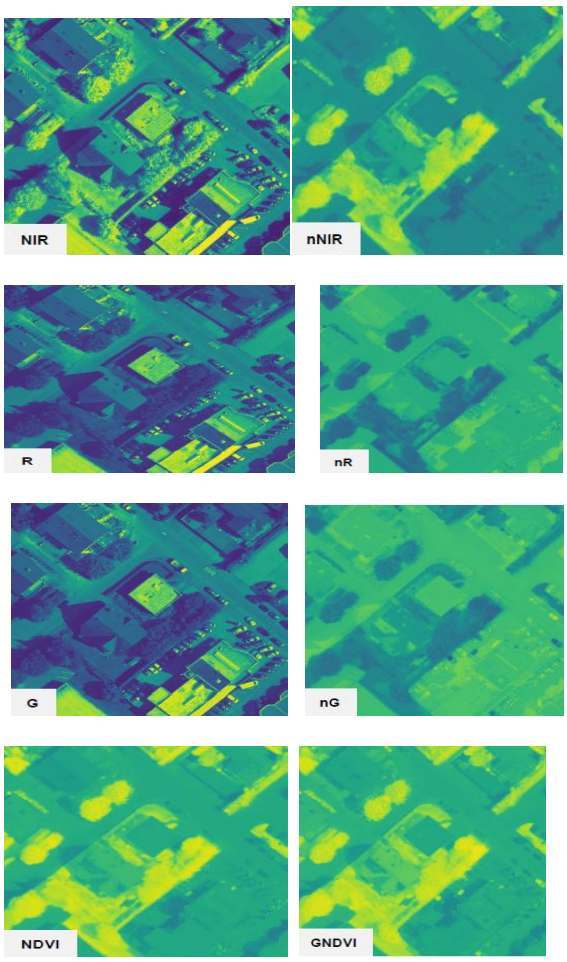

Figure 1. Visualization of the used Radiometric Information

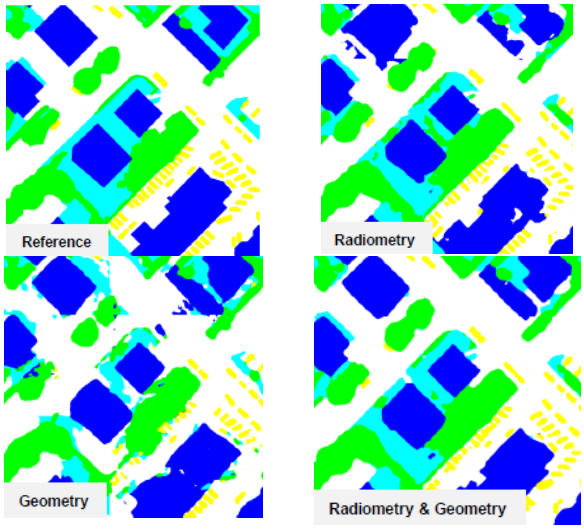

Figure 2. Comparision of semantic segmentation between only Radimetric, only Geometric and both features

V.CONCLUSION

In this paper, we have focused on the use of multi-modal data for the semantic segmentation of aerial imagery. Using true orthophotos, the corresponding DSMs and further representations derived from both of them, we have defined different sets of feature maps as input to a deep network.For the latter, we have proposed a Residual Shuffling Convolutional Neural Network (RSCNN) which combines the characteristics of a Residual Network with the advantages of atrous convolution and a shuffling operator to achieve a dense semantic labeling. Via performance evaluation on a benchmark dataset, we have analyzed the value of radiometric and geometric features when used separately and in different combinations for the semantic segmentation task.

The derived results clearly reveal that true orthophotos are better suited as the basis for classification than the DSM, the nDSM and different representations of geometric information and their combination. However, the combination of both radiometric and geometric features yields an improvement of the classification results. The derived results also indicate that some features such asthe NDVI are less suitable, and that the use of many features as the basis for semantic segmentation can decrease the predictive accuracy of the network and might thus suffer from the Hughes phenomenon. We conclude that selected data representations derived via feature extraction and feature selection techniques provide a gain if used as the basis for deep semantic segmentation.

REFERENCES

[1]. Audebert, N., Le Saux, B. and Lef`evre, S., 2016. Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks. In: Proceedings of the 13th Asian Conference on Computer Vision (ACCV), Taipei, Taiwan, Vol. I, pp. 180–196.

[2]. Audebert, N., Le Saux, B. and Lef`evre, S., 2017. Beyond RGB: very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing. In Press.

[3]. Badrinarayanan, V., Kendall, A. and Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12), pp. 2481–2495.

[4]. Blomley, R. and Weinmann, M., 2017. Using multi-scale features for the 3D semantic labeling of airborne laser scanning data. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Wuhan, China, Vol. IV-2/W4, pp. 43–50.

[5]. Chen, K., Fu, K., Gao, X., Yan, M., Sun, X. and Zhang, H., 2017. Building extraction from remote sensing images with deep learning in a supervised manner. In: Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, pp. 1672–1675.

[6]. Chen, K., Fu, K., Yan, M., Gao, X., Sun, X. and Wei, X., 2018. Semantic segmentation of aerial images with shuffling convolutional neural networks. IEEE Geoscience and Remote Sensing Letters 15(2), pp. 173– 177.

[7]. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L., 2016. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv preprint arXiv:1606.00915.

[8]. Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C. and Zhang, Z., 2015. MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.1274.

[9]. Cramer, M., 2010. The DGPF-test on digital airborne camera evaluation – Overview and test design. PFG Photogrammetrie – Fernerkundung – Geoinformation 2/2010, pp. 73–82.

[10]. Demantk´e, J., Mallet, C., David, N. and Vallet, B., 2011. Dimensionality based scale selection in 3D lidar point clouds. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Calgary, Canada, Vol. XXXVIII-5/W12, pp. 97–102.

[11]. Gerke, M., 2014. Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen). Technical report, ITC, University of Twente.

[12]. Volpi, M. and Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Transactions

on Geoscience and Remote Sensing 55(2), pp. 881–893.

[13]. Weinmann, M., 2016. Reconstruction and analysis of 3D scenes – From irregularly distributed 3D points to object classes. Springer, Cham, Switzerland.

[14]. Weinmann, M. and Weinmann, M., 2018. Geospatial computer vision based on multi-modal data – How valuable is shape information for the extraction of semantic information? Remote Sensing 10(2), pp. 1–20.

[15]. Weinmann, M., Schmidt, A., Mallet, C., Hinz, S., Rottensteiner, F. And Jutzi, B., 2015. Contextual classification of point cloud data by exploiting individual 3D neighborhoods. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Munich, Germany, Vol. II-3/W4, pp. 271–278.

[16]. West, K. F., Webb, B. N., Lersch, J. R., Pothier, S., Triscari, J. M. And Iverson, A. E., 2004. Context-driven automated target detection in 3-D data. Proceedings of SPIE 5426, pp. 133–143.